DOCUMENT RESUME

ED 442 843 TM 031 263

AUTHOR Loomis, Susan Cooper
TITLE Research Study of the 1998 Civics NAEP Achievement Levels.
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE 2000-04-27
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
CONTRACT ZA97001001
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Academic Standards; *Civics; Classification; Elementary Secondary Education; *National Competency Tests; *Teachers; *Validity
IDENTIFIERS *National Assessment of Educational Progress; *Standard Setting

ABSTRACT

In 1999 a validation study of the 1998 Civics Assessment of the National Assessment of Educational Progress (NAEP) was conducted, combining a similarity classification study and a booklet classification study. The rationale was that if the teachers who had participated in the 5-day achievement levels-setting process could not use the descriptions to judge student performance, it is unlikely that anyone could. If their classifications were very different from the performances in the test booklets, it would suggest that the cutpoints did not denote performance consistent with the achievement level descriptions (ALDs). Eleven eighth-grade teachers from the pilot study of achievement level setting panel participated. Teacher panelists tended to classify their own students higher than their performance levels on the special form of the NAEP developed for this study. When the same teachers were asked to classify the performance of students represented in the special Civics NAEP test booklets, they tended to classify those at or below the empirical score classification of the students' performance. These findings suggest that standards set with a booklet classification method will be higher than those set with the item-by-item method used for the NAEP achievement level setting process. Even teachers well trained in the NAEP achievement levels tend to overestimate the knowledge and skills of their students with respect to the ALDs. Overall, however, results provide information needed to confirm that the general achievement levels-setting process appeared to "work" in that panelists were able to use the ALDs in a different setting and for different purposes and the "translations" with respect to the score scale seem reasonably on target. An appendix contains classification forms from the similarities classification study, a list of participants in the validation portion of the study, and an agenda for the study session. (Contains 14 references.) (SLD)

Research Study of the 1998 Civics NAEP Achievement Levels

by

Susan Cooper Loomis
NAEP ALS Project Director
ACT, Inc.

This paper was prepared for presentation at the Annual Meeting of the American Education Research Association, April 27, 2000, New Orleans.

The research for this paper was supported by contract ZA97001001 with the National Assessment Governing Board.

Research Study of the 1998 Civics NAEP Achievement Levels

Susan Cooper Loomis[1]
NAEP ALS Project Director
ACT, Inc.

Background of the Study

In 1995 ACT designed and conducted two validation studies related to the achievement levels-setting (ALS) process for geography and U.S. history (ACT, 1996). NAGB requested that ACT implement similar studies as a part of the 1998 NAEP ALS validation studies. The decision was made to conduct the study for civics only.

The 1995 studies were called the Similarities Classification Study (SCS) and the Booklet Classification Study (BCS). The 1999 validation study for civics combined the two studies into a single study. This combination of the two studies provided the opportunity to address some questions that were raised in the 1995 studies. In particular, the findings from the Similarities Classification Study showed that teachers classified their own students at the same level or one achievement level higher than the student's actual or empirical performance level on NAEP. That finding would indicate that the achievement levels had been set too high. A contrary result was found for the Booklet Classification Study. Panelists classified the performance level represented by booklets as being at the same or one level lower than the empirical score level classification relative to the cutscores. This finding would indicate that the achievement levels were set too low. It was not clear whether these were general findings or findings due to differences in panel members for the two studies. The same panel members participate in both types of classifications (SCS and BCS) for the 1998 Civics NAEP study.

The logic of the Similarities Classification Study (SCS) is to test whether teachers who participated in the ALS process are able to apply the achievement levels descriptions (ALDs) in a way that is consistent with their use of the descriptions in setting the achievement levels. In particular, teachers were asked to make judgments about the knowledge and skills of their own students and the performance of their own students relative to the achievement levels descriptions. Those judgments were to classify the civics knowledge and skills of each student with respect to the achievement levels descriptions and to classify the expected

---

[1] Several people contributed significantly to the research project described in this paper. Dr. Robert Brennan and the members of the Technical Advisory Committee on Standard Setting (TACSS) largely developed the original design for the study conducted in 1995. Dr. Luz Bay helped to facilitate those meetings and to prepare the reports on the research. Those analyses provided guidance to analyses for this study.

Staff at ETS, National Computer Systems (NCS), Westat, Aspen Systems, the National Assessment Governing Board, and the National Center for Educational Statistics helped in many ways to assure the success of this project.

For the 1999 study, Wen-Hung Chen, Wen-Ling Yang, Jim Sconing, Brad Hanson, and Teri Fisher all played a significant role in the design, implementation, and analysis of results of the study. Paul Nichols and Patricia Hanick both helped with data analysis and evaluation of results after the study was conducted. Members of both ACT's Technical Advisory Team (TAT) and the TACSS reviewed earlier drafts of this report on the study.

I wish to acknowledge the assistance of Teri Fisher and Paul Nichols in preparing this report.

The final report on the study has not yet been completed for distribution. Persons interested in having a copy of that report when it is ready are urged to contact the author.

performance of each student on a special 100-minute form of the Civics NAEP with respect to the achievement levels descriptions.

These teachers are well trained in the ALDs during the ALS process. As a result of that experience, they should have a very good understanding of the meaning of the ALDs. If these teachers could use the ALDs to estimate achievement and performances of their students in a way consistent with the students' performance, then this would add support to the levels set. If not, then it seems unlikely that people who are not well trained in the ALDs and NAEP matters would be able to make reasonable interpretations about the meaning of the achievement levels. The teachers who have been ALS panel members provide the best-case scenario. If they are unable to use the ALDs consistently, then one must expect misinterpretations regarding the outcomes of the NAEP ALS process.

The Booklet Classification Study (BCS) design has been implemented a number of times by ACT. It was used in validation research studies in 1995 for geography and US history and again in 1997 for science. The design was tested as a method for setting achievement levels for writing in 1998. In previous studies, however, the booklets have been NAEP booklets of students who were assessed for the regular assessment period of 50 minutes. The number of items included in those NAEP booklets was not sufficient to provide a reliable score to represent performance of an individual student. For the 1999 validation study in civics, a special form of NAEP was developed that would produce a reliable estimate of individual student performance. These booklets were used for the BCS portion of the study, and teachers were asked to classify booklets into one of the three achievement levels, or at the Below Basic level.

Findings from the 1995 SCS indicated that the achievement levels were perhaps set too high. That is, teachers in both geography and US history tended to classify the knowledge and skills of their students at a level higher than the empirical performance level (i.e., higher than the achievement level category of the students' scores on the special form of NAEP), and the same was true for the level at which they classified the expected performance of their students on the special form of NAEP.

Findings from the 1995 SCS were countered by findings from the 1995 BCS. The Booklet Classification Study for both geography and US history indicated that the achievement levels were perhaps set too low. That is, panelists, including teachers, nonteacher educators, and representatives of the general public for each subject, generally classified performance represented in student NAEP booklets at one achievement level lower than the empirical performance level indicated by the score for the booklet.

Whether that was a general finding related to an inherent difference in the judgment tasks or a difference due to other factors was not known. For example, we were unable to determine whether the two sets of panelists differed to an extent that would have accounted for the differences in the classifications for the two types of tasks. A further complication with the BCS design was that the booklets used were actual NAEP booklets selected on the basis of plausible values. NAEP does not include enough items to produce a reliable individual student score. There was skepticism about using these booklets to represent individual student performances. Raw scores on the NAEP booklets tend to correspond to lower levels of achievement than is the case for the plausible values associated with the performance level for the booklet (ACT, 1997). That difference could have accounted for the judgments indicating a relatively lower level of performance.

The current study design included the same panelists to classify the expected performance of their students *and* to classify the student booklets. Further, the special form of NAEP designed for the study included enough items to provide a reliable individual score estimate. Further design features in the selection of booklets helped to eliminate the effect of the NAEP policy of treating "not reached" items as "not administered." Teacher panelists, in particular, have great difficulty in trying to take this policy into consideration when classifying test booklets according to achievement level categories.

This study not only informs the question of the validity of the Civics NAEP ALS process, but it also informs the debate regarding the use of item-by-item rating methods for the NAEP ALS process.

The Assessment Form Used for this Study

A special form of the grade 8 Civics NAEP was developed. This form included items selected to maximize representation of the entire grade 8 item pool with respect to item difficulty, content (5 areas of civics instruction), and type of items (multiple choice and constructed response). In addition, the information functions for the selected items were examined, relative to the entire item pool. Four blocks were selected to use in the study. All items remained in the same item order for each block. Please see Chen (1999) for a description of the criteria used and the analysis of item blocks selected for the study.

Two booklets were printed for each student including two cognitive blocks in each booklet. The second booklet for each student included the three sections to collect information from students on background demographic data, courses taken and topics studied in civics and social studies, and study behavior and test motivation.

Two different forms were developed for administration, and both forms included exactly the same items and sections. Form A included item blocks arranged in ascending numerical order and Form B included item blocks arranged in descending numerical order. This design allows for examination of a fatigue effect. Students were given a 10-15 minute break between the two booklets, i.e., after about 50 minutes of assessment time.

In addition to administration of the special form of the NAEP, ACT also collected information on the School Questionnaire and the Teacher Questionnaire for each school and teacher included in the study. We collected this information in order to be able to explore possible sources of differences, should there be sizable differences between the performance of these students and that of the national sample.

Logistics

Panelists had been told during the pilot study and the ALS that there would be a validation study involving grade 8 teachers. In January 1999 ACT contacted panelists and schools to secure their agreement to participate.

Our goal was to assure that every detail of this assessment and administration was as similar to a "standard" NAEP as possible. We provided schools with a choice of letters to be used to secure parental permission for students to participate in the study. Copies of the form letters prepared by ETS for NAEP distribution were obtained for the 1995 study, and those were again used in this study. ACT selected the items, secured copyright permission for materials to be included in the test booklets, and designed the cover for the new test booklets. NCES had requested that the cover be clearly distinguishable from the 1998 Civics NAEP. NCS printed the booklets, packaged spiraled sets of them for each administration session scheduled, and shipped them to Westat field staff. Westat administrators contacted the schools and established the exact schedule for testing, as well as the location for administration. They worked out all remaining details with the schools, and they contacted ACT regarding any "nonstandard" requests or circumstances. Westat staff returned booklets to NCS for scoring. NCS provided ACT with preliminary data about 10 days prior to the provision of the final, quality controlled data files.

Test Administration

Westat prepared to administer the assessment during the first two weeks of May to students in 13 schools. The schools included in the study were those represented by grade 8 teachers who served on the panel for either the pilot study or the ALS process for the 1998 Civics NAEP. Twenty-eight teachers were eligible to participate in the study: 11 teachers in the grade 8 pilot study panel and 17 in the ALS panel. Only 13 had been able to get both agreement from their schools to have their students participate in the study and to

meet the planned schedule for the study for administration and for the validation study panel meeting. Indeed, the panelists were originally asked to participate in the study scheduled for March, but that schedule had to be changed. The beginning of May was a feasible time for Westat administrators to help with the study, and 13 schools and teachers agreed to participate in the study at that time.

The clearance from OMB was delayed as a result of several factors. This caused a last-minute change in schedules for administration. Westat contacted each school and scheduled new administration dates a few days later. Three of the schools could not accommodate the change in schedule. Some of them already had other assessments scheduled, and some were too near the end of their school year. Two of the schools *included* in the assessment ended their school year in May: May 25, and May 26 were the closing dates for those schools. The remaining schools ended the school year between June 4 and June 22.

Westat prepared instructional materials for the administrators and shared those with ACT for review and approval. In a letter to teachers, ACT explained that the study design would be jeopardized if they saw the NAEP to be administered to their students. Teachers were also asked to avoid discussions with students about the test difficulty, their performance, and the particular items on the assessment. Administrators were given similar instructions. Westat administrators reported no unusual circumstances or problems during the administration.

Participation in the Study

Teacher Panelists

Eighth grade civics teachers who had served on either the pilot study or ALS panel were invited to participate in the validation study. A total of 28 teachers had participated in the process. Only 13 teachers were able to participate and to get agreement from their schools to have their students participate in the study. Due to delays in receiving the OMB Clearance and to the fact that the school term was ending very soon for some schools, only 11 schools were included in testing.

Teachers whose students could not be included in the assessment because of the delays and changes in scheduled test dates were invited to participate in the panel study. A total of 11 teachers participated as panel members for the study. Panelists included two teachers whose students were not tested; one teacher whose students were tested did not participate. Two of the teachers had been pilot study panelists, and the remaining 9 had been ALS panelists. The largest number of students tested for any one teacher was 57 and the smallest was 31. Only two study panel members were males.

Students

The schools participating in the study are located in the northeast (3), southeast (5), and central (3) NAEP regions. A list of schools and the number of students assessed is included in the Appendix to this paper.

A total of 499 student records were produced. Thirty-six records had no response data, so they were excluded. In addition, one student did not return for the second half of the assessment and another was reported to have been absent for a significant portion of the test period. Both of those student records were excluded. Therefore, the total valid number of student records from the assessment is 461.[2]

As reported above, one teacher did not participate in the classification study although her students were assessed. There were 30 students in that class, and those students were excluded from this report in order to have consistency in the numbers of students included in the different classifications reported for this study.

[2] For details about the number of records, number of items, coding of missing data (omits and not reached), and an assortment of other topics, please see Yang (1999).

In addition, the performance of 17 students was extremely low. The theta estimates for their performance was very low—lower than $\theta = -6$. The Technical Advisory Committee on Standard Setting (TACSS) recommended that we eliminate students with theta estimates lower than –3.0 because score estimates in that performance region are very unstable. The final count of students to be included in this study report is 414.

The racial/ethnic identity of these students was somewhat different from that of the national sample. (Please see Table 1.) In particular, there were fewer white and African American students and more Hispanic students.

Table 1
Racial/Ethnic Identity of Students in the National NAEP Sample
Compared to the Special Study Sample

Racial/Ethnic ID	National Sample %	Study Sample % (n=414)
White	67.1%	57.9%
Black	14.5	10.2
Hispanic	13.7	21.7
Asian/Pacific Islanders	3.4	5.6
American Indian/Alaska Natives	1.2	0.7

Coupled with the relatively higher racial/ethnic diversity for the students in this study (relative to the national sample for grade 8 in the Civics NAEP), a lower percentage of the students in this study are life-long U.S. residents (85% of the students in the study vs. 91% for the national sample) and a larger percentage are rather recent arrivals to this country (2.7% vs. 1% have lived in the US less than 3 years). A slightly higher percentage of the students in the study reported that a language *other than* English is spoken in the home all or most of the time (17% vs. 12%), and nearly a third of them reported that a language other than English is spoken in the home at least half of the time.

The educational attainment of the parents of these students is not very different from that of the national sample. About 43% of the students reported that their mothers had graduated from college, and 36% of them reported that their fathers had completed college. These figures for the national sample of students are 38% and 36%, respectively.

In terms of the civics topics studied, more students in these classes reported having studied the different topics than was the case for students in the national sample. Over 90% of these students reported having studied the US constitution and Congress, compared to only 78.5% of the national sample. Over 80% of these students reported studying the legislative process, the court system, political parties and the electoral process, and state and local governments, whereas only about 67% of the students in the national sample studied these topics. About 80% of these students reported having studied the President and Executive branch, compared to about 54% of the national sample. Finally, less than half of these students reported having studied governments of other countries and international relations/organizations, compared to about one-third of the national sample.

Descriptive statistics for the performance of students in each school are reported below in Table 2. The ACT NAEP-like scale values[3] computed as the mean and median performance levels for the schools are

[3] The ACT NAEP-like scale is a linear transformation of the NAEP scale. The values in this table were computed directly from scores on the theta metric. ACT uses the ACT NAEP-like scale scores in all ALS processes so that security of data is maintained.

reported, along with the numbers of students included in the computations and the minimum and maximum score values. The achievement level category associated with the mean and median scores is also noted. Both the mean and median performances of students in eight of the nine schools were in the Basic range of achievement. The mean and median scores for students tested in only one school were in the Proficient range. Only three schools had students scoring in the Advanced score range (178 and above), while eight schools had students with scores below the Basic level (below 149). Only one school, the school with the mean and median performance levels in the Proficient range (165 – 177.9), had no student scoring below the Basic achievement level, all others had at least one student scoring at the Below Basic level.

Table 2
Performance on the Special Form of the Civics NAEP
by Students Tested Whose Teacher Participated in the Classification Study

Mean Scores on the ACT NAEP-Like Scale for Students in this Report (n=414)	Median Scores on the ACT NAEP-Like Scale for Students in this Report (n=414)	N Students in Report from the School	Minimum Score	Maximum Score
161.02 (B)*	161.72 (B)	45	136.80	176.00
161.44 (B)	163.26 (B)	40	134.56	176.98
164.38 (B)	164.52 (B)	37	145.62	175.58
151.92 (B)	152.06 (B)	51	118.04	169.00
162.63 (B)	165.50 (B)	43	127.42	182.86
174.18 (P)	173.06 (P)	62	158.50	184.68
150.24 (B)	152.06 (B)	35	116.36	169.00
155.00 (B)	158.78 (B)	59	121.82	179.92
157.03 (B)	156.96 (B)	42	129.10	173.06

*B = score was within range of Basic cutscores. P = score was within range of Proficient cutscores. No school had a mean or median score at or above the Advanced cutscore.

Seventh Grade Class

Please note that one school included in the special assessment included mostly seventh grade students. The fact that the teacher had almost no eighth-grade civics students was revealed at no time during the ALS process or during the validation study. This fact was revealed when ACT staff read the comments of this teacher regarding the reasons for which students were classified at specific levels of achievement. The teacher's comments revealed that 35 of the 37 students included in the assessment were 7th grade students with little civics instruction. He also indicated that seven of those students were repeating 7th grade—one for the third time. In addition, two 8th grade students in the study were reported to be repeating the social studies course for the third time.

The mean performance level for the 7th grade students was the lowest of any of the 9 schools reported (150). The median performance level was tied with one other school (152). The performance of the students from this school was analyzed relative to that of all other students in the study. In general, the students appeared to be low performing students. Written comments by the teacher indicated that 15 students had very low grades in all classes, and two of the students were getting straight F's. The teacher also reported that five of the students were "borderline Special Ed" students. One of the five actually qualified for Special Education Classes, but his parents refused the placement.

Aside from low performance, the major difference between these students and other students in the study seems to be that these students did not study topics in civics. It is a bit difficult to understand how some students in a course reported that they had studied a topic and some reported that they had not. Whatever

the explanation, students' responses to questions included in the Civics NAEP booklet show that smaller proportions of these students reported having studied the various topics in civics and larger proportions responded to the questions with "Don't Know." These responses indicate that the students were, perhaps, such low ability students that they actually did not know what they had studied.

Teachers were not required to teach a civics course, per se, to participate on the Civics ALS panels; but the guidelines for participation on a panel require that the teacher panelists teach a social studies course for which civics content is a significant component. That guideline does not appear to have been met in this case. This teacher commented that a lack of civics would hold four [of his/her better] students back on the test.

The Study

The eleven teachers were convened for the two-day study in St. Louis. The study was conducted in the same place that had been the meeting site for the Civics ALS process. A meeting agenda is included in the Appendix to this paper.

Panelists arrived on the day before the study was officially begun. The meeting began at 8:30 AM on the first day; adjournment was scheduled for 5:00 PM on the second day. In fact, panelists completed their work early, and the meeting adjourned by about 3:00 PM on the second day.

Panelists in the ALS process are assigned to table groups of 5 persons each. The table assignments from the ALS were reviewed to assure that panelists were assigned to different table groups for this study. Three persons were assigned to three table groups and two to another. Civics panelists had been found to be rather difficult, and the new assignments were made in an effort to minimize the recurrence of belligerent behavior. In fact, the panelists convened for this study were among the most cooperative, supportive, and pleasant of any participants in a NAEP ALS study.

Panelists were first given an overview of the validation study: the design, the purpose, and the importance of conducting validation studies. They were given an opportunity to ask questions about the study. Next, they were given an overview of the ALS process in which they had participated. Data from Round 3 and the final cutpoints resulting from the achievement levels-setting process were shared with panelists. A brief description of subsequent studies and analyses that had been conducted was presented to the panelists. The study facilitator made it clear that ACT had analyzed many, many aspects of the ALS process and that the results of those analyses had been shared with the Technical Advisory Committee on Standard Setting. The facilitator also made it clear to the panelists that the National Assessment Governing Board had accepted the recommendations from ACT and that ACT had recommended the results of the ALS process.

The same content expert who had worked with the panelists in the ALS process retrained them in the Framework and the ALDs. Exercises to help panelists become "recalibrated" with respect to performance levels and achievement levels descriptions were also included as training for each step in the process. Panelists were given a set of exemplar items and student papers to review as part of their recalibration process.

Panelists were asked to complete four process questionnaires administered at different times throughout the process.

Implementation of the Similarities Classification Study

Classification of the overall level of civics knowledge and skill. The first part of the study is called the Similarities Classification Study (SCS), and it includes two different classifications. Teachers were asked to give classifications for each of their students included in the classes that had participated in the special

NAEP assessment. The name of each student in the assessment was printed on a classification form for each teacher. (Please refer to the Appendix for an example of the form.)

Teachers were first asked to classify the overall achievement of civics knowledge and skills for each student, according to the grade 8 achievement levels descriptions. Panelists were instructed to mark the location on the form that best corresponds to each student's level of knowledge and skills in civics, relative to the achievement levels descriptions. They were instructed to base their classifications on the achievement levels descriptions. Teachers were also asked to rate their confidence in their judgment of the achievement level classification of each student. Confidence ratings were simply *low, medium*, or *high*.

Teachers had been told to bring their grade books to the meeting or to review them carefully before coming. They were instructed in the use of their grade books and cautioned not to confuse course grades with performance relative to the achievement levels. They were to mark their classification of each student on the scale printed on the form for each student. A total or ten locations were identified for marking: solid Below Basic, upper Below Basic, lower Basic, solid Basic, upper Basic, and so forth through solid Advanced (no upper Advanced). In addition to marking the location to represent student performance relative to the achievement levels, panelists were also asked to mark their level of confidence (high, medium, or low) regarding the achievement level classification for each student. Teachers performed this task much faster than anticipated. All had finished within about one hour. Panelists completed the first process evaluation questionnaire at the end of that session.

An on-site change in the study design: Round 2 of classifying the overall level of civics knowledge and skills. During the first classification of their students, teachers commented on how their students would perform on NAEP items and some other aspects of student test-taking behavior. The facilitator cautioned them each time to focus only on the ALDs and their judgment of each student's knowledge and skills in civics.

Observers from the NAGB staff recommended a change in the study design to collect more information. They suggested that a sort of replication study be conducted before the second planned classification of students. This replication was to determine whether the panelists changed their classifications of students and to collect information for each classification regarding factors that potentially influenced teacher's classifications.[4]

Half of the students (every other student on the roster) for each teacher were included in the study. A copy of the classification form is included in the Appendix to this report. Teachers were asked to rate each of six factors that "could have influenced your classification of this student." They were given a Likert-type scale (5 =*very large influence*; 3 = *some influence*; 1 = *no influence*) for responding with regard to the following factors:

1. Overall knowledge and skills in all subjects
2. Overall knowledge and skills in civics
3. Test-taking behavior
4. Achievement levels descriptions
5. Items on the Civics NAEP
6. Grade(s) in my course

[4] NAGB staff worked on the list of factors and the format of the questions, but they ultimately left this task to the study facilitator and coordinator. Student files were transmitted electronically to the study site, and new classification forms were developed for the second classification of student achievement with respect to their overall civics knowledge and skills.

The panelists were told that this extra classification was added for purposes of collecting more information regarding the classification process. This addition was, in part, made because the amount of time required for the classifications was considerably less than had been scheduled. The panelists did not appear to be bothered by this additional task. This second round of the first classification procedure required almost twice as much time as the first. A copy of this form is included in the Appendix to this paper.

During the process of classifying half of their students in the second round of judging *overall civics knowledge and skills*, teachers commented that having the list of factors was causing them to think about those factors now, although they had not taken them into account during the first round. The results of this classification, therefore, seem contaminated, and ACT recommends that little attention be given to these classifications.

Analyses were conducted to determine whether these factors appeared to have influenced comments collected in the second, planned classification in the study. There was no evidence that this was the case. Perhaps the difference in the two classification tasks in the SCS accounts for the fact that there was no apparent influence from these factors on the classifications collected for students' expected performance on the special form of NAEP. That is, the factors influencing the teachers' classifications of the overall knowledge and skills of their students appear to be quite different from the more individualized factors influencing classifications of expected student performance on the special form of NAEP.

Classification of expected student performance on the special Civics NAEP. Before beginning the second type of classification task in the SCS, teachers were engaged in training exercises. These exercises included review and discussion of ten student booklets that had been pre-classified and identified as being within one of the three achievement levels categories or at the Below Basic level. Each panelist was also given a copy of all items in the grade 8 Civics NAEP, along with the scoring rubrics for each. In addition, they were given their own Reckase Charts from Round 2 of the ALS process in which they had been panelists. (Please see Loomis, 2000a and 2000b for more information about Reckase Charts.) The Reckase Charts allowed them to review their own ratings for specific items relative to the grade level cutscores and relative to their own cutscores for each achievement level. The charts also provide them with information regarding the relative difficulty of items in the item pool for grade 8. The purpose of these exercises was to help panelists become re-calibrated with respect to the cutscores set for each achievement level. These reviews helped panelists focus on the level of item difficulty in the assessment and how that level, along with the contents of the items, related to the achievement levels descriptions.

Teachers were instructed in the criteria used to select items for the special form of NAEP. They were told that the items met these criteria and that the assessment of cognitive items lasted 100 minutes. They were also told that students had a break after 50 minutes of testing. They were not told which items were included on the special assessment. They were instructed to again rate their confidence in their judgment of the achievement level classification of each student. Finally, they were instructed to comment on the factor(s) taken into account in classifying the expected performance of each student.[5] A copy of this classification form is included in the Appendix to this paper.

Implementation of the Booklet Classification Study

General Description. For the second part of the study, teachers were asked to use the ALDs as the criterion for classifying performances represented by student test booklets. This is called the Booklet Classification Study (BCS). This study was conducted in 1995 for U.S. history and geography and in 1997 for science. Although both the SCS and the BCS were conducted in 1995 for history and geography, different panelists were used for the two sets of studies in 1995. The fact that the same panelists were included in both the SCS and BCS for civics was judged to be a positive and significant change in the study design. Panelists

[5] Copies of instructions to panelists are available from the author upon request.

for previous booklet classification studies had included general public representatives and educators who were not K-12 classroom teachers, as well as teachers who met the same criteria as the teacher panelists included in this and other ALS studies.

Training for this procedure began after lunch on the second day. Teachers were told that there would be 40 booklets in the Booklet Classification Study (BCS). They were told only that booklets were selected such that the score of at least one booklet fell within the range of each achievement level. The scores of the booklets were not revealed to the teachers, nor were the empirical achievement level classifications. Individual item scores were not revealed to panelists. Panelists had scoring rubrics for all items, and they could refer to those rubrics.

The forty test booklets included in the classification study were carefully selected according to criteria recommended by the Technical Advisory Committee on Standard Setting (TACSS). Namely, the distribution of booklets, relative to the empirical score levels associated with each achievement level, was as follows: 7 scored below Basic, 13 scored within the Basic range, 13 scored within the Proficient range, and 7 scored within the Advanced range. The booklets were selected so that they were neither right around the cutscores nor all clustered at the midpoint. TACSS suggested this in order to eliminate booklets with scores that would likely be ambiguous with respect to the empirical scores for classification. A more complete description of the method of selecting booklets is available in Yang (1999c).

A total of 461 students were identified that met the criteria. The number of booklets scored within each achievement level range was counted. The total number of booklets within an achievement level category was divided by the target number of booklets to be included in the study for that level. This computation indicated the proportion of booklets to include in order to select the target number. Every n^{th} booklet was selected, beginning with the first n-count.

Some substitutions were made in order to include a student booklet from as many different schools as possible at each level. No school had an especially large number of students included in the sample. Further, booklet substitutions within a school were made in order to balance the number of student booklets from the two forms included in the BCS.

TACSS also suggested that booklets with a substantial number (4-5) of items left blank at the end of any block should be excluded, if possible. This was not difficult because few items were coded as "not reached."

Panelists were given a form to use for marking the achievement level category in which each booklet was classified. Four achievement level categories were available for the panelists to select. Booklets were numbered from 1-40, and the numbering was unrelated to the score of the booklet and unrelated to the school identification.

The practice session of the Booklet Classification Study. Panelists were instructed in the method and in marking their classification forms. In particular, they were told that they were to classify the booklets according to the achievement levels descriptions and to base their classifications on a holistic judgment. The facilitator stressed that scoring booklets was not the task and that booklet scores were not necessary in order to perform the task.

Panelists were given 10 booklets to classify in a practice session. They were given one hour for this practice classification. They were told that the rate (10 booklets per hour) would be the approximate rate necessary for the actual BCS the following day. All panelists completed classification of all booklets in the practice set within the allotted time. Several just managed to complete the task in that time, and most seemed somewhat concerned with the pace that would be required for the task the following day.

Following the classification, panelists were given the opportunity to discuss their classifications. This discussion was a whole group discussion so that all panelists would hear all comments. Panelists gained a sense of how their classification judgments compared to others in the group. They reported that this discussion helped them feel more confident about their preparation for the task the following day. (Please refer to Hanick, 1999 for more detailed information regarding the panelists' evaluations of the practice session.)

Findings of the Study

Similarities Classification

Data Reporting. Data reported in the tables of results are averages of the percentages of students classified in each achievement level category by each individual teacher. Separate cross tabulation tables were first formed for each teacher. The teacher's classification of the students and the empirical score classification (achievement level range associated with the students' NAEP scores) were entered on the table for each teacher. The cells of those tables were averaged to form the overall tables used in reporting data for this report.

A "hit rate" was computed for each table. This value is the percentage of classifications by teachers that correspond to the student score classifications into achievement level categories (empirical score classification). P_A reports the hit rate, and it is the sum of the percentages of students for which the teacher classifications correspond to the empirical score classification. P_E reports the "expected" value of the hit rate. That is, P_E reports the expected value of the hit rate, assuming that the distribution of classifications by teachers and the distribution according to the empirical scores classified by cutscores as independent classifications.

In addition to reporting the hit rate (P_A) and the expected hit rate (P_E), Cohen's Kappa statistic is also reported. This value is the proportion of "hits," correcting for the expected number of hits. Thus,

$$K = (P_A - P_E)/(1 - P_E)$$

shows the method of computing the Kappa statistic. Cohen's Kappa is +1 if the correspondence is exact, and it is 0 if there is no correlation between the two classifications. The greater the correspondence between classifications, the closer the K will be to +1.

Findings for the Classification of Student's Overall Civics Knowledge and Skills. Results reported in Table 3 indicate that the teachers were somewhat more likely to classify their students' level of overall civics achievement and expected performance on the special NAEP at a higher level than the student's empirical performance. Findings in the 1995 study indicated the same pattern. That is, teachers tend to think that their students' achievement and performance are as high or higher than the actual performance of their students on the special form of NAEP. The comparisons here are between the level at which teachers classified their students *overall knowledge and skills in civics* and the level at which the student scores were classified, i.e., the empirical classification.

Table 3
Percentage of Students Classified within Achievement Level Categories Based on Overall Civics Knowledge and Skills Relative to the Empirical (MLE) Score Classifications

Table N = 414	Achievement Level Classification of Overall Civics Knowledge and Skills (SCS#1)			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	**Below Basic** (n=72)	**Basic** (n=118)	**Proficient** (n=119)	**Advanced** (n=105)
Below Basic (<149.2) (n=64)	**8.7%** (n=36)	5.6% (n=23)	1.2% (n=5)	0.0 (n=0)
Basic (149.2 – 165.39) (n=189)	8.2 (n=34)	**19.1** (n=79)	15.0 (n=62)	3.4 (n=14)
Proficient (165.4 – 177.89) (n=140)	0.5 (n=2)	3.9 (n=16)	**11.8** (n=49)	17.6 (n=73)
Advanced (≥ 177.9) (n=21)	0.0 (n=0)	0.0 (n=0)	0.7 (n=3)	**4.4** (n=18)

Bold entries are for cells that would represent "hits" or agreement.

P_A=.440
P_E =.267
K = .243

Teachers' classifications of the *overall civics knowledge and skills* of their students agreed with the students' empirical performance level in 44% of the cases.[6] The association between score estimates of the achievement level performance categories of students and teachers' classifications of the overall knowledge and skills of the students is positive, but somewhat low (K = .24). Teachers tend to estimate the civics knowledge and skill achievement level category of their students to be higher than that derived for students on the basis of their performance on the special form of NAEP. This is evident in the relatively larger percentages in the cells above the diagonal of values in Table 3 compared to the percentages in the cells below the diagonal in Table 3. Teachers' classifications of overall civics knowledge and skills were within one achievement level of the empirical level for 95% of the students, and they were the same or one level higher in 87% of the cases. Teachers classified the overall civics knowledge and skills of their students at a lower level than the empirical achievement level for only 13% of the students. This finding is similar to that for geography and US history in 1995.

Of the 64 students who scored in the Below Basic category, teachers classified 36 of them in that category. This means that 56% of the students scoring below the Basic cutscore were also classified at that level by their teachers who rated their overall knowledge and skills in civics at the Below Basic level.

Of course only half (36) of the students classified at the Below Basic level (72) by their teachers actually scored at that level.

Of 21 students who actually scored at or above the Advanced level, teachers classified 18 of them as Advanced with respect to their overall knowledge and skill level in civics. There is no higher level at which students could be classified. That upper limit, coupled with the tendency for teachers to classify their

[6] The forms used for teachers' classifications included 10 locations for marking achievement, described above. In order to compare results to the empirical score classifications and to the results of the Booklet Classification Study, the classifications were coded into only 4 categories: Below Basic, Basic, Proficient, and Advanced.

students at a higher level, makes this higher correspondence (86%) rather unexceptional. Even more important is the fact that teachers classified 105 students in the Advanced category with respect to their overall knowledge and skills in civics. Seventy percent of those students actually scored in the Proficient score range, and 13% scored in the Basic level.

Teachers' judgments of students corresponded with the student performance scores for about 42% of the students (79 of 189) scoring in the Basic achievement level range. Of the students (118) that teachers judged to be within the Basic level with respect to their overall knowledge and skills in civics, about two-thirds (79) actually scored within that range of performance. This compares to only about 41% of the students that teachers judged to be within the Proficient level (with respect to overall knowledge and skills in civics) that actually scored at that level (49 of 119).

Achievement levels descriptions were used to set the cutscores that served to classify student scores on the special form of the NAEP. Teachers who participated in that process used these descriptions to classify the overall knowledge and skills of their students in civics. Perhaps the relatively low correspondence between the two classifications is a result of the fact that teachers were asked to judge "overall" knowledge and skills in civics. Factors other than those included in the achievement levels descriptions are likely to be taken into account in making this judgment about specific students that the teachers know personally. The achievement levels descriptions served to filter these factors when teachers were asked to characterize their students' overall knowledge and skills with respect to the achievement levels, but the two factors being compared here (overall knowledge and skills *versus* performance on a special form of the Civics NAEP) are rather distant. That distance was greatly decreased in the second set of judgments teachers were asked to make in the Similarities Classification Study.

Findings for the Classification of Students' Expected Performance on the Special Form of NAEP. For the second set of classifications, teachers were asked to classify each student according to the achievement level category that would be expected for the student's performance on the special form of the Civics NAEP. Teachers were familiar with all the items in the grade 8 NAEP item pool, but they did not know which specific items were included on the special form used to assess their students. They knew the length of the assessment and the circumstances of the administration.

A pattern of findings similar to that reported in Table 3 is found in Table 4 for the relationship between teachers' expectations of the performance of their students on the special form of NAEP and the actual performance of their students on the assessment. The correspondence between actual performance on the special form and teachers' estimates of student performance was expected to be somewhat higher than that for estimates of overall knowledge and skills. Although teachers did not know the exact contents of the assessment, it seemed likely that the classification of *expected performance* would be more similar to actual student performance than the classification of overall knowledge and skills. Teachers' classifications of the *expected performance of their students on the special form of NAEP* agreed with the empirical performance level in about 44% of the cases. The value of the Kappa statistic is approximately the same as that for Table 3, and this again shows a low association between the two classifications.

Table 4
Percentage of Students Classified within Achievement Level Categories Based on Expected Performance on the Special Form of the Civics NAEP Relative to the Empirical (MLE) Score Classifications

Table N = 414	**Achievement Level Classification of Expected Student Performance on the Special Form of the Civics NAEP (SCS#2)**			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	**Below Basic** (n=87)	**Basic** (n=115)	**Proficient** (n=119)	**Advanced** (n=93)
Below Basic (<149.2) (n=64)	**9.7%** (n=40)	5.3% (n=22)	0.5% (n=2)	0.0 (n=0)
Basic (149.2 – 165.39) (n=189)	10.9 (n=45)	**17.9** (n=74)	15.0 (n=62)	1.9 (n=8)
Proficient (165.4 – 177.89) (n=140)	0.5 (n=2)	4.6 (n=19)	**12.6** **(n=52)**	16.2 (n=67)
Advanced (≥ 177.9) (n=21)	0.0 (n=0)	0.0 (n=0)	0.7 (n=3)	**4.4** (n=18)

Bold entries are for cells that would represent "hits" or agreement.

P_A=.446
P_E =.268
K = .243

Teacher's classifications of the *expected performance* of their students were within one achievement level of the students' empirical performance level in 97% of the cases. Overall, teachers classified the expected performance of 39% of their students at a higher achievement level than the empirical performance level, and they classified the expected performance on only 17% of their students as lower than the empirical level. Teachers classified expected student performance at a higher level than the empirical classification almost as frequently as they classified the performance in the same category as the empirical score classification. Please see Table 4 for these results comparing the teachers' classifications of the *expected performance of their students on the special form of the NAEP* and the classifications of the students' actual performance, based on the MLE estimates of their scores and the NAEP cutscores for each level.

Table 5 provides summary data for the results of the classifications by teachers of the *overall civics knowledge and skills of their students* (SCS#1), by teachers of the *expected performance of their students on the special form of NAEP* (SCS#2), and summary data of the cutscore classifications based on the MLE score estimates (empirical performance scores levels). The percentages of students classified at each achievement level category by teachers are very similar for the two classifications teachers made. There were more students classified at the below Basic level and fewer at the Advanced level, when expected performance on the special form of the Civics NAEP was the frame of reference than when overall knowledge and skills was used for the classification. Relative to the empirical score classifications, the proportions classified by both sets of teacher judgments in the Below Basic level and the Advanced level look quite high. Relative to the empirical score classifications, the proportions classified in the Basic and Proficient levels look quite low.

Table 5

Percentage of Students Classified within Achievement Level Categories Based on Overall Civics Knowledge and Skills (SCS#1) and Expected Performance on the Special NAEP (SCS#2) Relative to Percentages Based on MLE Scores Within Cutscore Ranges for each Achievement Level (Empirical Score Level)

	SCS #1 % Classified Within Level (n=414)	SCS #2 % Classified Within Level (n=414)	% Within Empirical Score Level (N=414)
Below Basic	17.4% (72)*	21.0% (87)	15.5% (64)
Basic	28.5 (118)	27.8 (115)	45.7 (189)
Proficient	28.7 (119)	28.7 (119)	33.8 (140)
Advanced	25.4 (105)	22.5 (93)	5.1 (21)

*Numbers in parentheses are the numbers of students classified in each achievement level category.

Table 6 reports the mean and median values of MLE empirical score estimates, reported on the ACT NAEP-like score metric, of students who were classified according to each of the classification criteria in Table 5. These data help to provide more understanding of the level of similarity in the performance of students classified for each classification condition.

The data in Table 6 show that the scores on the special form of NAEP were considerably higher for students classified by their teachers at the Below Basic level than the scores that were classified according to the achievement level cutscores. Teachers generally classified their students at a higher level than the empirical score classification level. The data reported in Table 6 seem to be generally consistent with that finding. Consistent with the data reported in Tables 3-5, the scores of students in the classification levels determined by teacher judgments were generally lower than those classified by empirical score levels. At the Below Basic level, however, the mean and median scores of students classified by their teachers are higher than those classified according to the empirical score level.

Table 7 reports the correspondence between teacher's classifications of students with respect to the student's overall knowledge and skills in civics and the student's expected performance on the special form of NAEP. Overall, teachers tended to classify the overall civics knowledge and skills of their students as highest, and their expected performance as next highest. The two classifications corresponded for 73% of the students, and the classification of overall civics knowledge and skills was higher than expected performance on the special form of NAEP for 18% of the students. For only 9% of the students did teachers classify the expected performance on the special NAEP as higher than the overall knowledge and skills in civics.

Table 6

Mean and Median ACT NAEP-Like Score Values for MLE Score Estimates of Performance by Students: Three Classifications of Performance

Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores) (Table n = 414)		Overall Knowledge and Skills	Expected Performance on Special Form of Civics NAEP	Empirical Score by Achievement Level
Below Basic (<149.2)	Mean	146.18	147.30	137.92
	Median	148.84	149.82	140.58
	n =	(72)	(87)	(64)
Basic (149.2 – 165.39)	Mean	154.72	155.42	157.52
	Median	156.40	156.82	157.8
	n =	(118)	(115)	(189)
Proficient (165.4 – 177.89)	Mean	163.54	163.96	170.68
	Median	164.24	164.94	170.4
	n =	(119)	(119)	(140)
Advanced (≥177.9)	Mean	172.08	173.20	181.18
	Median	172.08	172.92	181.18
	n =	(105)	(93)	(21)

*Numbers in parentheses are the numbers of students classified in each category.

Teachers' classifications of their students with respect to overall civics knowledge and skills were more similar to their classifications with respect to expected performance on the special form of NAEP than either of these classifications was to the empirical performance classifications.

Agreement between the first round of classifications based on *overall civics knowledge and skills* and the second round (added on site), including only half of the students, was quite high: 84% of the students were classified in the same achievement level in the two rounds. Four percent of the students were classified at a higher level the second time than the first, and 12% were classified at a lower level the second time than the first. Given the suspected impact of the factors on the second round form, the agreement seems remarkably high between the two rounds of classifying students' overall civics knowledge and skills. Data in Table 8 show the correspondence in classifications of students' overall civics knowledge and skills in the two rounds.

Table 7

Percentage of Students Classified within Achievement Level Categories Based on Overall Civics Knowledge and Skills (SCS#1) by Expected Performance on the Special NAEP (SCS#2)

Table N = 414	**Achievement Level Classification of Overall Civics Knowledge and Skills (SCS#1)**			
Achievement Level Classification of Expected Performance on Special Form of Civics NAEP (SCS#2)	**Below Basic** (n=72)	**Basic** (n=118)	**Proficient** (n=119)	**Advanced** (n=105)
Below Basic (n=87)	**14.0%** (n=58)	6.5% (n=27)	0.5% (n=2)	0.0 (n=0)
Basic (n=115)	3.4 (n=14)	**18.6** (n=77)	5.6 (n=23)	0.2 (n=1)
Proficient (n=119)	0.0 (n=0)	3.4 (n=14)	**20.1** (n=83)	5.3 (n=22)
Advanced (n=93)	0.0 (n=0)	0.0 (n=0)	2.7 (n=11)	**19.8** (n=82)

Bold entries are for cells that would represent "hits" or agreement.

P_A=.725
P_E =.255
K = .631

Table 8

Percentage of Students Classified within Achievement Level Categories based on Overall Civics Knowledge and Skills, Rounds 1 and 2 of SCS#1

Table N = 211	**Achievement Level Classification by teachers of Overall Civics Knowledge and Skills: Round 1 (SCS 1, #1)**			
Achievement Level Classification by teachers of Overall Civics Knowledge and Skills: Round 2 (SCS 1, #2)	**Below Basic** (n=41)	**Basic** (n=61)	**Proficient** (n=56)	**Advanced** (n=53)
Below Basic (<149.2) (n=51)	**19.0%** (n=40)	4.7% (n=10)	0.5% (n=1)	0.0 (n=0)
Basic (149.2 – 165.39) (n=56)	0.5 (n=1)	**22.8** (n=48)	3.3 (n=7)	0.0 (n=0)
Proficient (165.4 – 177.89) (n=53)	0.0 (n=0)	1.4 (n=3)	**20.4** (n=43)	3.3 (n=7)
Advanced (≥ 177.9) (n=51)	0.0 (n=0)	0.0 (n=0)	2.4 (n=5)	**21.8** (n=46)

Bold entries are for cells that would represent "hits" or agreement.

P_A=.840
P_E =.251
K = .786

Booklet Classification Study

For the Booklet Classifications, teachers were asked to classify 40 booklets into NAEP achievement level categories using the achievement levels descriptions as the criterion. The data reported in Table 9 show the correspondence between teachers' classification of booklets and the empirical score classifications that were based on MLE estimates of student performances. In this table, the data reported are for classifications by 11 teachers of 40 booklets selected from the empirical score classifications to produce this distribution of booklets: 7 Below Basic; 13 Basic, 13 Proficient, and 7 Advanced booklets.

Table 9

Correspondence of Teachers' Classifications of Student Booklets into Achievement Level Categories and Empirical Score Classifications of Student Booklets into Achievement Level Categories

Table N =440	**Achievement level classification of student booklets by teachers**			
Achievement level classification by empirical scores of student booklets (ACT NAEP-Like Cutscores)	**Below Basic** (n=137)	**Basic** (n=155)	**Proficient** (n=104)	**Advanced** (n=44)
Below Basic (<149.2) (n=77)	**15.9%** (n=71)	1.4% (n=6)	0.0 (n=0)	0.0 (n=0)
Basic (149.2 – 165.39) (n=143)	14.3 (n=63)	**17.7** (n=78)	0.5 (n=2)	0.0 (n=0)
Proficient (165.4 – 177.89) (n=143)	0.6 (n=3)	16.2 (n=71)	**14.1** (n=62)	1.6 (n=7)
Advanced (≥ 177.9) (n=77)	0.0 (n=0)	0.0 (n=0)	9.1 (n=40)	**8.4** (n=37)

Bold entries are for cells that would represent "hits."

P_A=.561
P_E =.263
K = .404

Panelists classified 56% of the booklets at the same achievement level as the empirical score classification based on MLE estimates for the booklets. Overall, they tended to classify the booklets at the same level as the empirical score level, or lower. The exact correspondence between teacher judgments and student performances was higher in this classification of booklets than for the previous two classifications involving estimates of performance for specific students. Classifications were within one achievement level of the empirical score classification for all except three of the booklets. Those three booklets were actually scored within the Proficient level, but they were classified as Below Basic. Only 9 booklets (2%) were judged to represent performance higher than the empirical score classification, and 177 booklets (40%) were judged to represent performance below the empirical score classification.

This pattern is exactly opposite that found in the two previous classifications. When teachers classified the expected performance *of their own students* and overall knowledge and skills *of their own students*, they classified student achievement at higher levels than student performance on the special form of NAEP would seem to warrant. When teachers classified actual performances of *anonymous* students on this special form of NAEP, however, their judgments led to classifications that were lower than the students' actual performance on the special form of NAEP would seem to warrant. This same pattern characterized the findings from the 1995 studies in U.S. history and geography (ACT, 1995). A booklet classification study for the 1996 Science NAEP also resulted in booklets classified at the same or lower levels of

achievement than the classifications based on empirical score estimates (ACT, 1997). Based on the consistent results of the booklet classification studies conducted by ACT, the decision would be to move the NAEP cutscores for these subjects and grades higher.

Table 10 reports the mean, minimum, and maximum ACT NAEP-like score values for performance scores of the booklets classified at each achievement level. These data allow more detailed evaluation of the relative levels of performances judged to represent different achievement levels. With one exception, a logical pattern of average scores is found for the booklets classified at each level. This means that teachers were able to discern a relative ordering of performance, even though they generally classified booklets at one level lower than the empirical score classification level. For example, the average score (143) of the booklets that were scored below the Basic cutscore but classified at the Basic level was higher than those scored below the Basic cutscore and classified at the Below Basic level (135). Similarly, the average score of booklets scored within the range of the Basic cutscores and classified at the Proficient level was 160, compared to those classified within the Basic range (159), and the Below Basic range (157). The exception was the average score of the three booklets that were classified in the Below Basic category but actually scored within the Proficient range. Those three booklets had an average score (171) that was about the same as the 62 booklets scored and classified within the Proficient range. At least one of the booklets judged to be in the Below Basic category was scored 174 on the ACT NAEP-like scale—well above the Proficient cutscore of 165.

Table 10
Mean, Minimum, and Maximum ACT NAEP-Like Score Values for
Performance Scores of Student Booklets Classified at Each Achievement Level

Table N =440	**Achievement level classification of student booklets by teachers**			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	**Below Basic** (n=137)	**Basic** (n=155)	**Proficient** (n=104)	**Advanced** (n=44)
Below Basic (<149.2) (n=77)	**134.8*** (119.3/146.3) (n=71)	**143.0** (136.7/146.3) (n=6)	**0.0** (n=0)	**0.0** (n=0)
Basic (149.2 – 165.39) (n=143)	**157.2** (153.6/163.7) (n=63)	**158.5** (153.6/163.7) (n=78)	**160.0** (158.5/161.6) (n=2)	**0.0** (n=0)
Proficient (165.4 – 177.89) (n=143)	**170.5** (169.0/173.8) (n=3)	**165.4** (169.0/175.6) (n=71)	**171.9** (169.0/175.6) (n=62)	**173.6** (172.4/175.6) (n=7)
Advanced (≥ 177.9) (n=77)	**0.0** (n=0)	**0.0** (n=0)	**181.6** (179.7/184.7) (n=40)	**182.2** (179.9/184.7) (n=37)

*Bold numbers are the mean ACT NAEP-like score value for the booklets classified at the level. Numbers within parentheses are minimum/maximum values of booklets classified at the level.

Comparisons of Each Teacher's Three Classifications of Their Own Students. The design for the civics study was changed somewhat from that used in 1995 in order to provide the opportunity to have the same panelists classify student booklets (the BCS) and their own students (the SCS). These study findings indicate that the cutscores are set too high when performance levels on NAEP are compared to teachers'

estimations of the achievement their students relative to the performance criteria. These study findings also indicate that the cutscores are set too low when performance levels on NAEP are compared to teachers' evaluations of students' actual performance on the assessment. There was a lack of certainty regarding the generality of these findings, however, because the differences could result from judgments by two different sets of panelists. Findings from this study have revealed the *same patterns of findings* produced in the two different studies for each subject in the 1994 ALS process. This adds credibility to the findings of both sets of studies.

A final set of comparisons for a limited number of students is possible, and these comparisons provide an even more direct test of whether these same patterns hold. The expected patterns are that teachers' classifications of their own students will be higher than the students' empirical performance relative to the NAEP achievement levels; and teachers' classifications of the performance of students whose identity is unknown to them will be lower than the students' empirical performance. That is, the same pattern found for teachers' classifications, overall, is expected for the classifications of the teachers' own students.

Thirty-seven of the 40 student booklets included in the BCS were assessments of students of the nine teachers in the study. During the BCS portion of the study, only one teacher commented that she could match the identity of the student to the booklet. Whether more teachers discerned the identity of the examinees is not known. It seems unlikely that this was a frequent occurrence since no one else commented on this during the BCS.

The data in Table 11 show the patterns found across the three classifications. Consistently, the data in Table 11 show that these teachers *tended* to classify the overall knowledge and skills and the expected performance on the special NAEP of their own students at the same level or at a higher achievement level than the level at which the students performed on the special NAEP. When the teachers read the responses of those same students, in order to classify the performance represented in the booklet, they *tended* to classify the booklet at a level lower than the level at which the student performed. Furthermore, teachers classified the examinee booklets of their students at lower levels than they classified the expected performance by the same students. Finally, teachers classified the examinee booklets of their students at lower levels of achievement than they classified the overall level of knowledge and skills for the same students.

It is particularly interesting to examine the correspondence between teachers' classifications of the test booklets of their students, i.e., how teachers classified the actual performance of their students, and their classifications of the expected performance of the same students on an assessment with the same attributes of the special form of NAEP that was administered to their students. The data in the lower right section of Table 11 show how teachers of the students in the study classified these two factors. Expectations of how the student would perform on such an assessment corresponded to the teacher's evaluation of the student's performance for only 16 of the 37 students (43%) for which all comparable data were available. Teachers classified the actual performance, i.e., the booklets, of 19 (51%) of the students at a lower level than the level at which the expected performance was classified. When booklet classifications are compared to the teacher's classification of the same student's overall knowledge and skills, the findings again show that teachers classified the booklets of 20 students (54%) at a lower level than they classified the overall civics knowledge and skills of the same students. Those data are reported in the middle section of the right side of Table 11. Teachers classified the performance of only 15 students (40%) at the same level as the overall knowledge and skills classification. In all, booklets for only two students were classified at a higher level of achievement than that judged to represent the student's overall knowledge and skill in the subject of the assessment.

Table 11

Relationships Between and Among Empirical Score Performance Classifications and Teacher Judgment Classifications of Performance by NAEP Achievement Levels

		Overall Knowledge & Skills Level				Expected NAEP Performance Level				Booklet Classification			
		BB	B	P	A	BB	B	P	A	BB	B	P	A
Empirical Classification	BB (6)	3	3	0	0	5	1	0	0	5	1	0	0
	B (12)	1	6	5	0	1	6	5	0	7	4	1	0
	P (12)	0	2	4	6	0	3	6	3	0	7	5	0
	A (7)	0	0	1	6	0	0	2	5	0	0	6	1
	Total	4	11	10	12	6	10	13	8	12	12	12	1
Overall Knowledge & Skills	BB (4)					4	0	0	0	5	1	0	0
	B (11)					2	9	0	0	7	4	1	0
	P (10)					0	1	9	0	0	7	5	0
	A (12)					0	0	4	8	0	0	6	1
	Total					6	10	13	8	12	12	12	1
Expected Performance	BB (6)									5	1	0	0
	B (10)									5	4	1	0
	P (13)									2	5	6	0
	A (8)									0	2	5	1
	Total									12	12	12	1

Summary

The general rationale behind this research design is that if teachers who participated in the five-day achievement levels-setting process *cannot* use the descriptions to judge student performance, it is unlikely that anyone can. So, if their classifications were wildly *different* from the performances in the booklets, we would tend to think that the cutpoints do not denote performance consistent with the ALDs.

Results of the study in civics show patterns similar to those found in 1995. That is, teacher panelists tend to classify their own students higher than their performance levels on the special form of NAEP developed for the study. This is true for classifications of the overall civics knowledge and skills of their students and for classifications of expected performance on the special form of the Civics NAEP developed for this study. The highest classification levels were for students' overall knowledge and skills in Civics, followed by classifications of students' expected performance on the special form of the Civics NAEP. Both of these classifications tended to be at or above the empirical score classification of the student's performance. When the same teachers were asked to classify the performance of the students, represented in the special Civics NAEP test booklets, they tended to classify those at or below the empirical score classification of the student's performance.

Discussion of Findings

This is the first booklet classification study conducted with reliable measures at the individual student level. Previous studies have indicated that the plausible values used in NAEP tend to increase performance levels, relative to the raw score classifications (ACT, 1997**the study with science booklets, etc. reported to TACSS and maybe in the final reports). This is ACT's first NAEP study with a BCS design for which the scores were not derived through plausible values. The findings of this study concur with the previous findings by ACT regarding standard setting with a booklet classification method. That is, the standards set with a booklet classification method will be higher than those set with the item-by-item method used for the NAEP ALS process.

Further, even when teachers are well-trained in the NAEP achievement levels descriptions and are familiar with the NAEP assessment pool, they are still likely to overestimate the knowledge and skills of their students with respect to the achievement levels descriptions of what students should know and be able to do, when those estimates are compared to the students' actual performance on the NAEP.

The study involved only grade 8 teacher panelists, so one cannot be certain that the same results would hold for teacher panelists at other grades. There is no reason to expect that the results for the grade 8 teachers would differ significantly from those for teachers at other grades, but data are not available to test that.

Further, one cannot generalize to a larger population of students with these results[7]. The study was designed to help focus on the ALS process and the ability of panelists to make rational and reasonable judgments. If the results indicate that these teachers would make significantly different judgments when using the ALDs with respect to expectations about their own students than they made in the rating process, then we would doubt the results of the ALS process. We would doubt that anyone could accurately interpret performance relative to the ALDs. On the other hand, if the results indicate that these teachers made judgments that are very similar to the judgments they made when setting the achievement levels, then we *cannot* say that others would make similar judgments.

One important feature of this study is that it combined both the SCS and the BCS designs. The 1995 studies in geography and US history showed that teachers tended to classify their own students at a higher level than the students' performance would warrant. We also found that panelists classified booklets of students,

[7] Paul Nichols presented research on the generalizability of the results in this session of AERA 2000. Please see Nichols, 2000.

unknown to them, at a level lower than the empirical score classification. There appeared to be a rather compelling logic to these patterns, but it was not clear that the findings were a result of differences in the sets of panelists or a more general behavioral judgment finding. Having the same panelists participate in both parts of the study provided the control on panelists to make direct comparisons across different classification tasks. Having teachers' classifications of the student's overall civics knowledge and skills, expected performance on the special form of the NAEP, *and* of their actual performance on the assessment provides data on all four classifications (including the empirical score classification into achievement level categories) by one teacher for one student.

The results of this study provide information needed to confirm that the general achievement levels-setting process for the 1998 Civics NAEP appeared to "work." That is, panelists were able to use the ALDs in a different setting and for different purposes, and the "translations" with respect to the score scale seem reasonably on target.

References

ACT (1995) *Research Studies on the Achievement Levels Set for the 1994 NAEP in Geography and U.S. History*. Iowa City, IA: Author.

ACT (1997). *Setting Achievement Levels on the 1996 NAEP in Science: Final Report, Volume IV: Validity Evidence and Special Studies*. Iowa City, IA: Author.

Chen, W.H. (1999). *Item Block Selection and Estimation of Reliability for the Similarities Classification Validation Study for the 1998 Civics NAEP*. Report to the Technical Advisory Committee on Standard Setting, April 29-30, 1999, St. Paul.

Hanick, P.L.(1999b). *1998 NAEP Civics Achievement Levels Validation Study: Panelists' Responses to Process Evaluation Questionnaires and Comments on Factors Influencing Classifications*. Report to the Technical Advisory Committee on Standard Setting, September 16-17, 1999, Atlanta

Hanick, P.L. (1999a). *Comparison of 1998 Civics NAEP Almanac Data and Background Questionnaire Data for Validation Study*. Report to the Technical Advisory Committee on Standard Setting, December 2-3, 1999, San Francisco.

Hanson, B. (1999). *Estimating the Reliability of Individual Scores for SCS*. Report to the Technical Advisory Committee on Standard Setting, February 18-19, 1999. Atlanta.

Loomis, S.C. (1999a). *1998 Civics NAEP Similarities Classification and Booklet Classification Study: Design Issues and Considerations*. Report to the Technical Advisory Committee on Standard Setting, April 29-30, 1999, St. Paul.

Loomis, S.C. (1999b). *1998 Civics NAEP Validation Study of Achievement Levels*. Report to the Technical Advisory Committee on Standard Setting, September 16-17, 1999, Atlanta.

National Academy of Education (1993a). *Setting performance standards for student achievement*. Stanford, CA: Author.

Nichols, P.D. (2000). *Generalizing Civics NAEP Achievement Levels to Teachers' Judgments*. Paper presented at the annual meeting of the American Education Research Association, New Orleans, 2000.

Sconing, James (1999). *Technical Notes for the Estimation of Likelihood Based Estimates of Student Ability for the 1999 NAE SCS*. Iowa City, IA: ACT.

Yang, W.L. (1999a). *Analysis of Fatigue Effects on Student Performance on the special Form of the Civics NAEP*. Report to the Technical Advisory Committee on Standard Setting, September 16-17, 1999, Atlanta.

Yang, W.L. (1999b). *Comparing MLE and EAP Estimation Outcomes for the 1999 Similarity Classification Study*. Report to the Technical Advisory Committee on Standard Setting, December 2-3, 1999, San Francisco.

Yang, W.L. (1999c). *Notes for Data Cleaning, File Management, and Scoring for the 1998 NAEP Civics Validation Study*. Report to the Technical Advisory Committee on Standard Setting, September 16-17, 1999, Atlanta.

Appendix

Similarities Classification Study

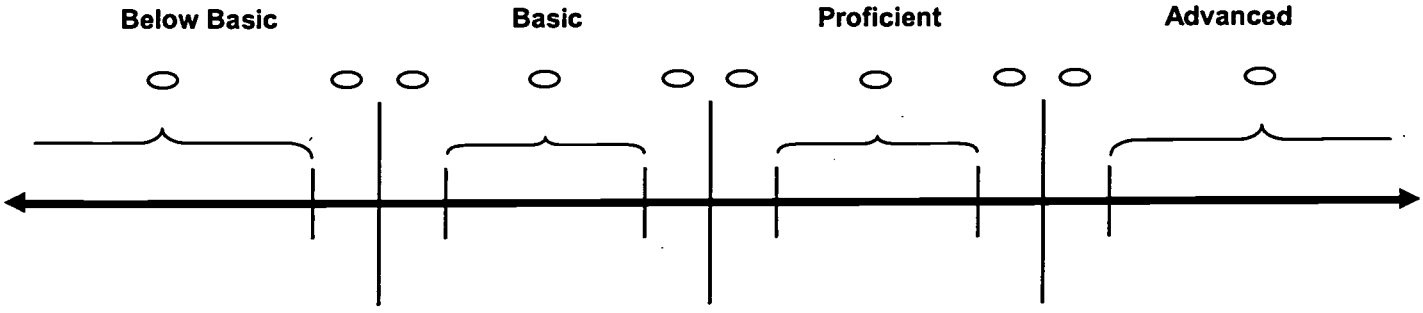
Classification 1

«Student_Name» _____ «Teacher_Name» _____

Student Teacher

○ Cannot classify (please explain): _____

_____



My level of confidence regarding this achievement level classification is (mark only one):

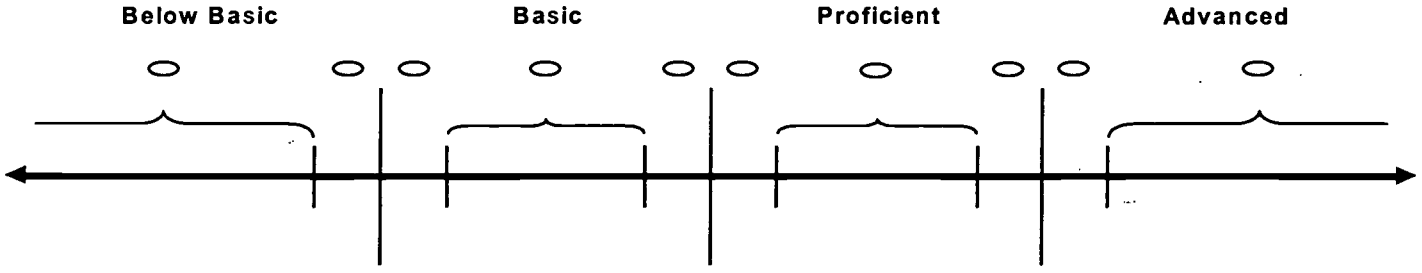
- ○ High
- ○ Medium
- ○ Low

«Student Name»
Student

«Teacher Name»
Teacher

○ Cannot classify (please explain): _____

_____



My level of confidence regarding this achievement level classification is (mark only one):

- ○ High
- ○ Medium
- ○ Low

Here is a list of factors that could have influenced your classification of this student. Please rate each factor.

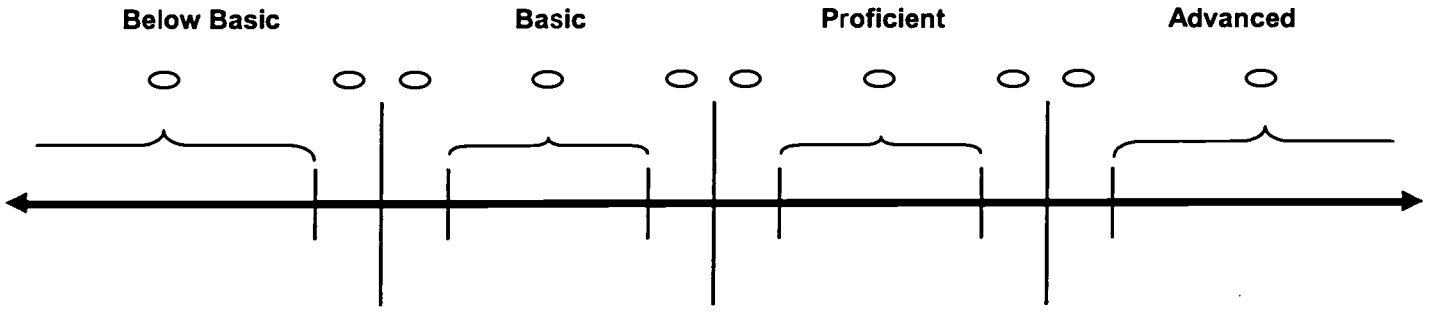
	Very Large Influence 5	4	Some Influence 3	2	No Influence 1
Overall knowledge and skills in all subjects					
Overall knowledge and skills in civics					
Test-taking behavior					
Achievement levels descriptions					
Items on the Civics NAEP					
Grade(s) in my course					

Similarities Classification Study

Classification 2

«Student_Name» — Student

«Teacher_Name» — Teacher



My level of confidence regarding this achievement level classification is (mark only one):

- ○ High
- ○ Medium
- ○ Low

Please comment on the types of things you took into consideration when classifying this student:

«Teacher_Name»

Teacher

Booklet Classification Form

	Below Basic	Basic	Proficient	Advanced
1.	O	O	O	O
2.	O	O	O	O
3.	O	O	O	O
4.	O	O	O	O
5.	O	O	O	O
6.	O	O	O	O
7.	O	O	O	O
8.	O	O	O	O
9.	O	O	O	O
10.	O	O	O	O
11.	O	O	O	O
12.	O	O	O	O
13.	O	O	O	O
14.	O	O	O	O
15.	O	O	O	O
16.	O	O	O	O
17.	O	O	O	O
18.	O	O	O	O
19.	O	O	O	O
20.	O	O	O	O

	Below Basic	Basic	Proficient	Advanced
21.	O	O	O	O
22.	O	O	O	O
23.	O	O	O	O
24.	O	O	O	O
25.	O	O	O	O
26.	O	O	O	O
27.	O	O	O	O
28.	O	O	O	O
29.	O	O	O	O
30.	O	O	O	O
31.	O	O	O	O
32.	O	O	O	O
33.	O	O	O	O
34.	O	O	O	O
35.	O	O	O	O
36.	O	O	O	O
37.	O	O	O	O
38.	O	O	O	O
39.	O	O	O	O
40.	O	O	O	O

Validation Study Participation

Teacher School	Estimated # Students	Total # Assessed
Peggy Allen *		
Greenville Junior High	29	0
Greenville, Illinois		
Marleni Burns		
Multicultural Magnet School	50	43
Bridgeport, Connecticut		
Mary Carter		
Alma Middle School	50	0
Alma, Arkansas		
Vicki Foley		
Campbell County Middle School	50	45
Alexandria, Kentucky		
Gretchen Gundlách		
EJD Middle School	50	42
Phoenix, New York		
Brian Gustafson		
Eisenhawer Middle School	60	41
Rockford, Illinois		
William Hardin		
Wayne Middle School	35	36
Wayne, West Virginia		
Leslie Lee		
Hammocks Middle School	70	57
Miami, Florida		
Denise Nickens		
Robinson Middle School	46	63
Wichita, Kansas		
Lynnette Poag		
Memorial Middle School	60	46
Vineland, New Jersey		
Susan Putnam		
Laurens Junior School	59	0
Laurens, South Carolina		
Janet Smith		
Kenneth Henderson Middle Schoc	60	60
Garden City, Kansas		
Jackie Viana *		
Hialeah Middle School	60	31
Hialeah, Florida		
* Did not participate in the study		

Civics NAEP Validation Research Study
Achievement Levels-Setting Process
Grade 8 Teacher Panelists

Ritz-Carlton Hotel, St. Louis
July 9-11, 1999

Agenda

Thursday, July 8

5:00 - 5:30 p.m.	*Amphitheater Prefunction* Check-in at ACT Registration Desk, 2nd Floor. Get name tag, final agenda, and information about transportation downtown tonight.

Friday, July 9

8:00 a.m.	*Consulate* Continental Breakfast
8:30 a.m.	Welcome: Susan Loomis (ACT) and Mary Lyn Bourque (NAGB) General Orientation Session: Susan Loomis • Overview of the study • Review of ALS process
10:00 a.m.	Break
10:15 a.m.	Review of Framework and Achievement Levels Descriptions (ALDs): John Patrick
11:15 a.m.	Exercises and Discussions to Re-train in Framework and ALDs • Review Eighth Grade Exemplar items
Noon	*Directors* Lunch
1:00 p.m.	Continue Re-training in Framework and Achievement Levels Descriptions *via* Exercises to practice use of ALDs
1:30 p.m.	Instructions for Estimating Student Civics Achievement: Classification #1
2:00 p.m.	Break
2:15 p.m.	Student Civics Achievement Classifications #1 Questionnaire #1
4:30 p.m.*	Adjournment

* This is an approximate time only. The actual amount of time required will depend upon the number of your students in the study.

Saturday, July 10

8:00 a.m.	*Consulate* Continental Breakfast
8:30 a.m.	Review Achievement Levels Descriptions and Framework *via* Discussions and Exercises • Student booklets • Grade 8 item pool
10:00 a.m.	Break
10:15 a.m.	Instructions for Estimating Student Civics Achievement: Classification #2
10:30 a.m.	Student Civics Achievement Classifications #2 * Questionnaire #2
Noon - 1:30 p.m.	*Promenade* Lunch Buffet
2:30 p.m.	The Booklet Classification Study • Instructions in Booklet Classification Process
2:45 p.m.	Practice Classification Session
3:45 a.m.	Break
4:00 p.m.	Discussion of Practice Classifications
4:45 p.m.	Questionnaire #3
5:00 p.m.	Adjourn

* You may break for lunch before you finish, if necessary. It is, of course, best to complete your classifications without a long break.

Sunday, July 11

8:30 a.m.	*Consulate* Continental Breakfast
9:00 a.m.	Booklet Classifications • Information about booklets included in the study • Instructions about classification procedures • Marking classification forms • Review Achievement Levels Descriptions
9:30 a.m.	Classify Booklets (breaks as needed)
Noon	*Colonnade* Lunch
1:00 p.m.	Re-calibration for Continuation of Booklet Classification Study • Questionnaire #4
4:00 p.m.	Wrap-Up Session: Questions and Answers
4:15 p.m.	Adjournment

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").